Storage Systems

OSPP Chap 12

Main Points

- File systems
  - Useful abstractions on top of physical devices
- Storage hardware characteristics
  - Disks and flash memory

File Systems

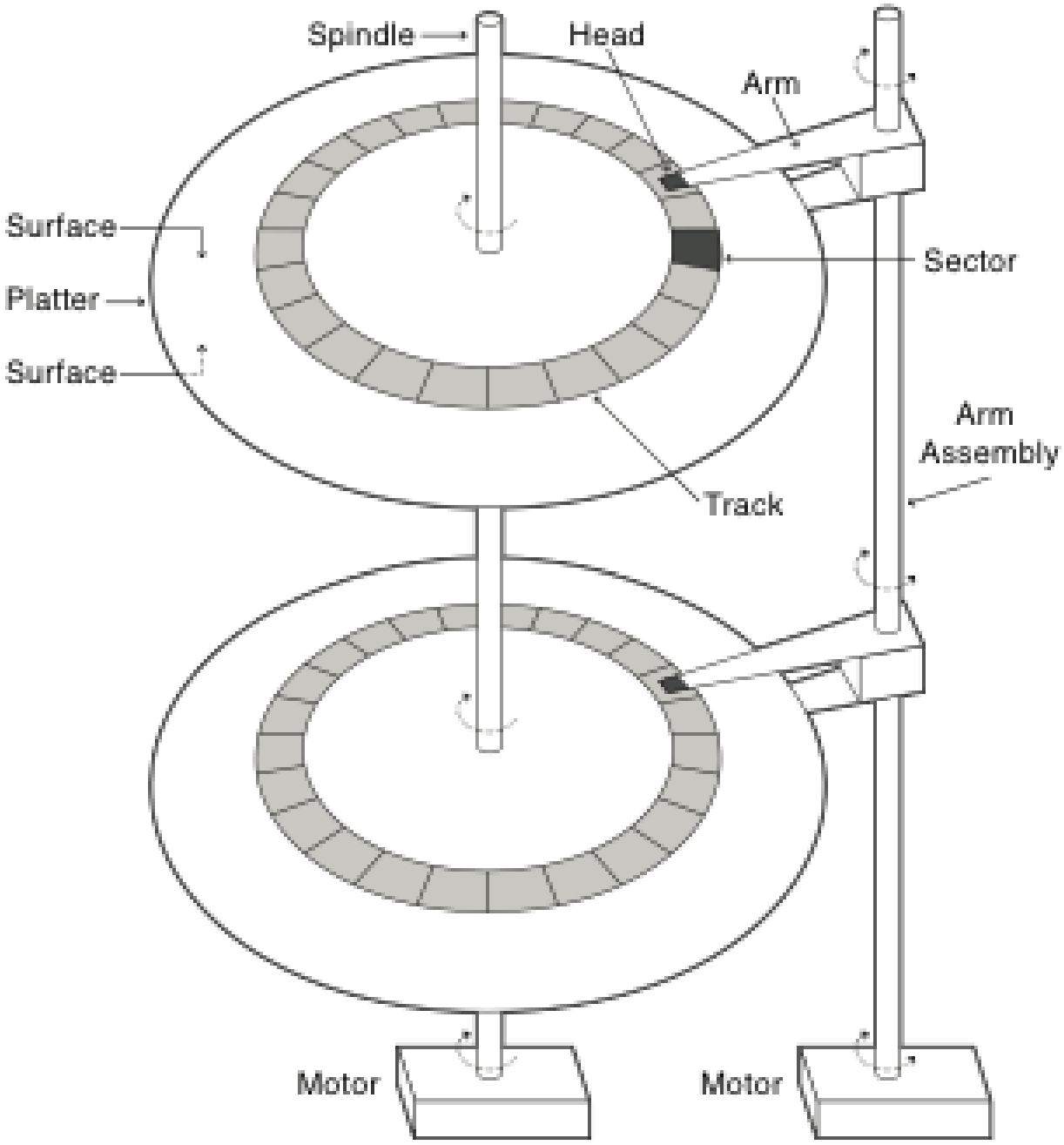
- Abstraction on top of persistent storage
  - Magnetic disk
  - Flash memory (e.g., USB thumb drive)
- Devices provide
  - Storage that (usually) survives across machine crashes
  - Block level (random) access
  - Large capacity at low cost
  - Relatively slow performance
    - Magnetic disk read takes 10-20M processor instructions

File System as Illusionist: Hide Limitations of Physical Storage

- Persistence of data stored in file system:
  - Even if crash happens during an update
  - Even if disk block becomes corrupted
  - Even if flash memory wears out
- Naming:
  - Named data instead of disk block numbers
  - Directories instead of flat storage
  - Byte addressable data even though devices are block-oriented
- Performance:
  - Cached data
  - Data placement and data structure organization
- Controlled access to shared data

Storage Devices

- Magnetic disks
  - Storage that rarely becomes corrupted
  - Large capacity at low cost
  - Block level random access
  - Slow performance for random access
  - Better performance for streaming access
- Flash memory
  - Storage that rarely becomes corrupted
  - Capacity at higher cost
  - Block level random access
  - Good performance for reads; worse for random writes


Spindle
Head
Arm
Surface
Platter
Surface
Sector
Arm
Assembly
Track
Motor
Motor

Sectors

Sectors contain sophisticated error correcting codes

- Hide corruptions due to neighboring track writes
- Read an entire sector

- Sector sparing
  - Remap bad sectors transparently to spare sectors on the same surface
- Track skewing
  - Sector numbers offset from one track to the next, to allow for disk head movement for sequential ops

Disk Performance

Disk Latency =

- Seek Time + Rotation Time + Transfer Time
- Seek Time: time to move disk arm over track (1-20ms)
  - Fine-grained position adjustment necessary for head to “settle”
- Rotation Time: time to wait for disk to rotate under disk head
  - Disk rotation: 4 – 15ms (depending on price of disk)
  - On average, only need to wait half a rotation
- Transfer Time: time to transfer data onto/off of disk
  - Disk head transfer rate: 50-100MB/s (5-10 usec/sector)
  - Host transfer rate dependent on I/O connector (USB, SATA, …)

Toshiba Disk (2008)

Size	
Platters/Heads	2/4
Capacity	320 GB
Performance	
Spindle speed	7200 RPM
Average seek time read/write	10.5 ms/ 12.0 ms
Maximum seek time	19 ms
Track-to-track seek time	1 ms
Transfer rate (surface to buffer)	54–128 MB/s
Transfer rate (buffer to host)	375 MB/s
Buffer memory	16 MB
Power	
Typical	16.35 W
Idle	11.68 W

Question

- How long to complete 500 random disk reads, in FIFO order?

Question

- How long to complete 500 sequential disk reads?

Disk Scheduling

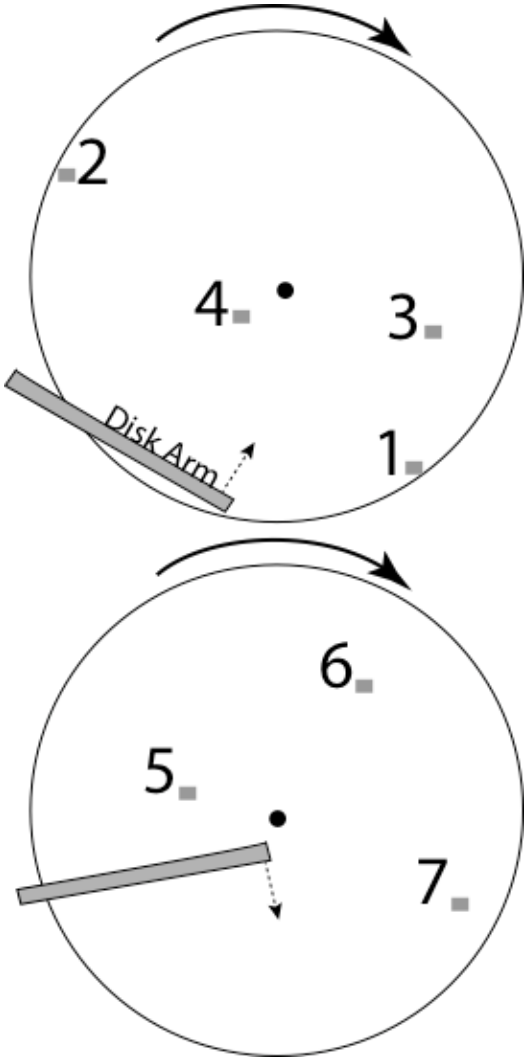
- FIFO
  - Schedule disk operations in order they arrive
  - Downsides?

Disk Scheduling

- Shortest seek time first
  - Not optimal!
    - Suppose cluster of requests at far end of disk
  - Downsides?

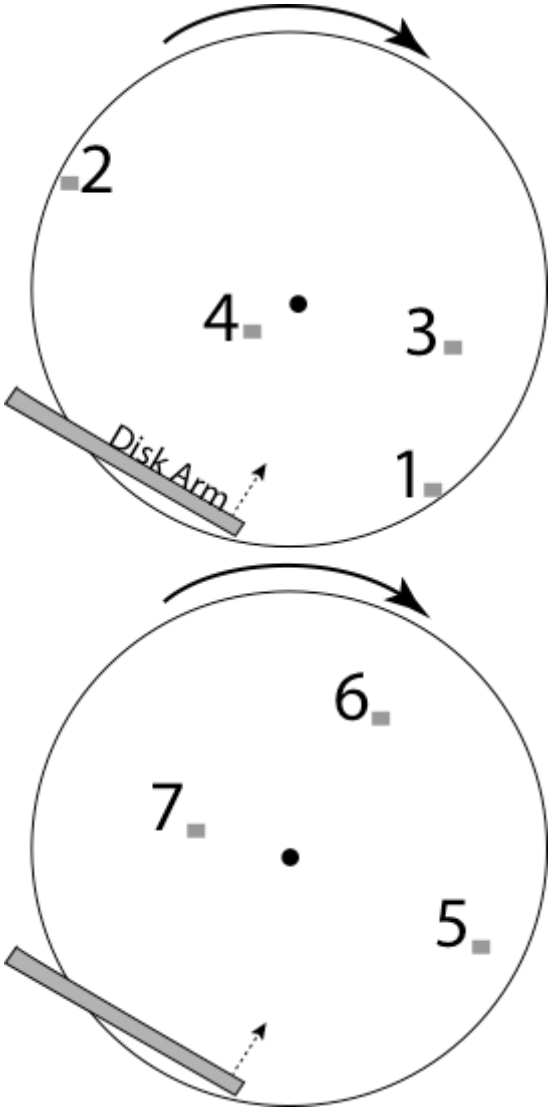
Disk Scheduling

- SCAN: move disk arm in one direction, until all requests satisfied, then reverse direction
- Also called “elevator scheduling”



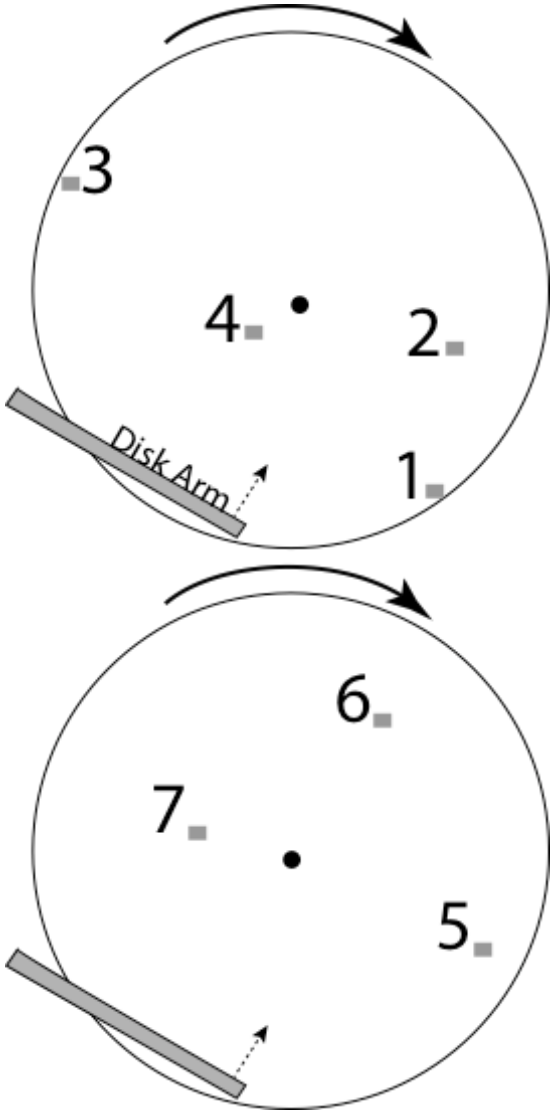
Disk Scheduling

- CSCAN: move disk arm in one direction, until all requests satisfied, then start again from farthest request



Disk Scheduling

- R-CSCAN: CSCAN but take into account that short track switch is < rotational delay



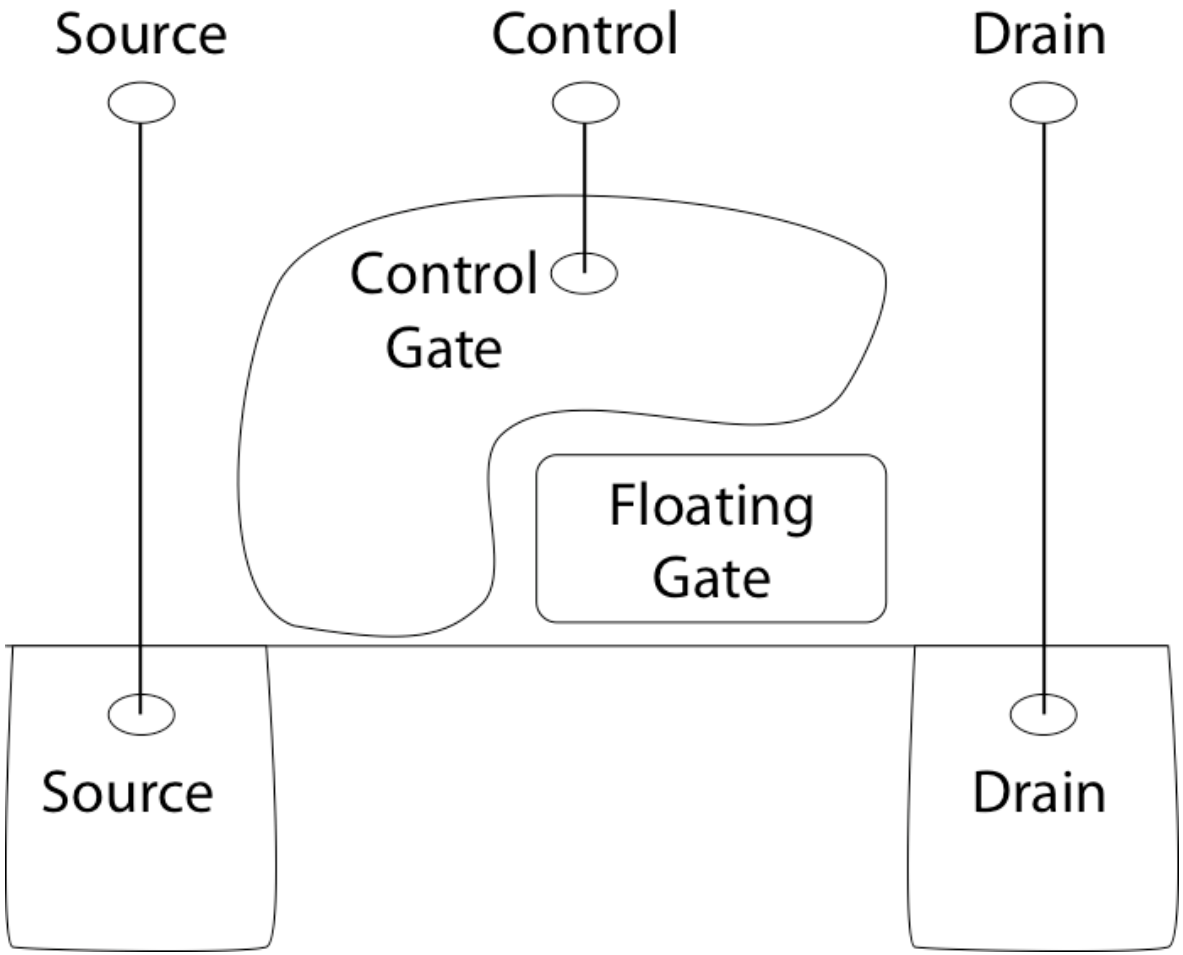
Question

- How long to complete 500 random disk reads, in any order?

Question

- How long to read all of the bytes off of a disk?

Flash Memory



Flash Memory

- Writes must be to “clean” cells; no update in place
  - Large block erasure required before write
  - Erasure block: 128 – 512 KB
  - Erasure time: Several milliseconds
- Write/read page (2-4KB)
  - 50-100 usec

Flash Drive (2011)

Size	
Capacity	300 GB
Page Size	4KB
Performance	
Bandwidth (Sequential Reads)	270 MB/s
Bandwidth (Sequential Writes)	210 MB/s
Read/Write Latency	75 μs
Random Reads Per Second	38,500
Random Writes Per Second	2,000 (2,400 with 20% space reserve)
Interface	SATA 3 Gb/s
Endurance	
Endurance	1.1 PB (1.5 PB with 20% space reserve)
Power	
Power Consumption Active/Idle	3.7 W / 0.7 W

Question

- Why are random writes so slow?
  - Random write: 2000/sec
  - Random read: 38500/sec

Flash Translation Layer

- Flash device firmware maps logical page # to a physical location
  - Garbage collect erasure block by copying live pages to new location, then erase
    - More efficient if blocks stored at same time are deleted at same time (e.g., keep blocks of a file together)
  - Wear-levelling: only write each physical page a limited number of times
  - Remap pages that no longer work (sector sparing)
- Transparent to the device user

File System – Flash

- How does Flash device know which blocks are live?
  - Live blocks must be remapped to a new location during erasure

Next Time

- Device drivers in Linux
- Read posted material
- Lab #4 (short) using device drivers